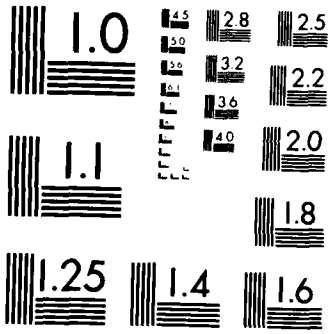
AD-A188 250 A SUMMARY OF PUBLISHED ARTICLES AND TECHNICAL REPORTS UNDER THE ONR CONTR..(U) ILLINOIS UNIV CHAMPAIGN K K TATSUOKA ET AL. SEP 87 N00014-82-K-0645 1/1

UNCLASSIFIED F/G 5/8 NL



(12)

A Summary of Published Articles and Technical Reports under the ONR Contract N00014-82-K-0645

APPENDIX TO FINAL REPORT

Kikumi K. Tatsuoka

Maurice M. Tatsuoka

Summary work between 1982-1987

September 1987

Introduction

The work done during this contract period was, in our opinion, very fruitful and significant in developing the foundations for a theory of cognitive error diagnosis. Below, we list the publications under four categories: I. Development of theory for rule space; II. Applications: learning, cognition, and measurement issues; III. Order analytic paper; and IV. Site curriculum development. Along with the list of categories there will be an abstract (or summary) and a discussion of each publication.

I. Development of Theory for Rule Space

1. Tatsuoka, K. K. (1983). Rule space: An approach for dealing with misconceptions based on item response theory. Journal of Educational Measurement, 20(4), 345-354 .

Abstract

This study introduces a new approach for representing large numbers of erroneous rules of operation found in a procedural domain of arithmetic. The model, based on item response theory, is capable of predicting the likelihood of each rule, and hence of remediating misconception(s) possessed by students. Moreover, this probabilistic approach helps to pinpoint weaknesses in instruction. The new model, called rule space, challenges the credibility of the traditional right-or-wrong scoring procedure because the same response patterns are sometimes produced by two different erroneous rules, and it is often the case that one incorrect answer is due to some serious misunderstanding of a basic concept while the other is due to a careless error.

Discussion

This paper introduced a new approach for handling large numbers of erroneous rules of operation quantitatively in signed-number subtraction and for investigating the likelihood of each erroneous rule. By taking this approach, remediation of students' difficulties and the modification of instructional materials will be focused on the real problems and on the parts of instruction that should be changed.

However, the rule-space concept raises several important issues in scoring procedures, analysis of response patterns, and their applications as well. First, the traditional binary scores are obtained by throwing away valuable information for the sake of simplification. Fortunately, the responses from signed-number addition and subtraction problems are decomposable into two component scores without losing any valuable information. (Note that 99% of the responses are obtained by adding or subtracting two absolute values while the signs of the answers may be either plus or minus.) Therefore, a pair of the two binary response patterns obtained by scoring each component separately could represent 99 percent of responses given by the students. All erroneous rules uniquely expressed by the two component response patterns of absolute value and sign parts (along with other responses not necessarily yielded by erroneous rules) are mapped into the rule space as well. However, this procedure may not work in some other cases. For example, the responses to addition and subtraction of two mixed fraction numbers may not be

decomposable into independent components.

A new scoring procedure by which all information on wrong options is used in the final form of scores is urgently needed, especially in the area of achievement of the problem-solving type. Computer programs that can diagnose students' misconceptions will be constructed in various fields in the future. This trend is inevitable as computers have penetrated into the areas of testing and teaching more deeply than ever, and will continue to do so in the future. As analysis of misconceptions advances further, the theory of educational measurement must be developed as well so that the misconception data (such as number of erroneous rules) can be properly handled without losing valuable information. The methods developed by Siegler and Anderson, Brown and his associates, and Tatsuoka and her associates use all the information contained in the students' responses for assessing rules. The component scoring procedure is an approach toward using all information contained in the responses. It is necessary to investigate methods comparable to component scoring for using more information provided by the students' responses in other areas in the future.

The second problem is that the same response patterns are sometimes produced by two different rules, and it is often the case that one is due to some serious misunderstanding of a basic concept while the other is due to a careless computational error. It must be possible to separate these two different sources of misconcpetions so that the information can be used in proper evaluation of teaching and instructional materials. Analyses of response patterns alone may not be a sufficient technique for analyzing students' misconceptions.

2. Tatsuoka, K. K., & Linn, R. L. (1983). Indices for detecting unusual patterns: Links between two general approaches and potential applications. Applied Psychological Measurement, 7(1), 81-96.

Abstract

Two distinct approaches, one based on item response theory and the other based on observed item responses and standard summary statistics, have been proposed to identify unusual response patterns of responses to test items. A link between these two approaches is provided by showing certain correspondences between Sato's S-P curve theory and item response theory. This link makes possible several extensions of Sato's caution index that take advantage of the results of item response theory. Several such indices are introduced and their use illustrated by application to a set of achievement test data. Two of the newly introduced extended indices were found to be very effective for purposes of identifying persons who consistently use an erroneous rule in attempting to solve signed-number arithmetic problems. The potential importance of this result is briefly discussed.

Discussion

As was shown above, the caution index which Sato (1975) developed, based solely on a comparison of observed item responses to group responses, can be readily extended to theory-based estimates of person and group response probabilities. The caution index is a linear transformation of the covariance of a person's response pattern with one or another theoretical curve computed using IRT. Alternatively, the extended caution indices may be viewed as linear transformations of the

covariance or correlation between a person's response pattern and a theoretical curve (either the PRC, as in the case of $ECE4_i$ and $ECI5_i$ or the GRC, as in the case of $ECI2_i$ or $ECI3_i$).

The application of the extended caution indices that were introduced in this paper provided strong evidence that the indices depending on the distance between a person's response pattern and his/her theoretical PRC (i.e., $ECI4_i$ and $ECI5_i$ or $ECI3_i$) are quite effective for purposes of identifying persons who consistently use an erroneous rule in answering signed-number arithmitic problems. This is a potentially important result that deserves further investigation with other data sets involving different types of achievement test data. If additional research yields similar results, these indices may have considerable instructional utility, because instruction can be made much more specific once it is determined that a student is consistently making an error as the result for a particular misconception.

3. Tatsuoka, K. K. (1984a). Caution indices based on item response theory. Psychometrika, 49(1), 95-110.

Abstract

A new family of indices was introduced earlier as a link between two approaches: One based on item response theory and the other on sample statistics. In this study, the statistical properties of these indices are investigated and then the relationships to Guttman Scales, and person response curves are discussed. Further, these indices are standardized, and an example of their potential usefulness for diagnosing students' misconceptions is shown.

Discussion

The extended caution indices, ECI1, ECI2, ECI4 and ECI6 are standardized by the usual tranformation,

$$SECIm = \frac{ECIm - E(ECIm \mid \theta_i)}{SE(ECIm \mid \theta_i)} \qquad \text{for } m = 1,2,4 \text{ and } 6.$$

The conditional expectation of ECI4 is a function of the θ level, but those of the other three ECIs are identically zero. If we sample two students from different T_i levels, then it is dangerous to compare their ECI4 values in order to determime which student's response patterns are more aberrant than the other's. Moreover, the standard errors of all three ECIs are functions of T_i and have "U" shaped trend curves. This explains the past findings that the correlation of personal indices, such as the caution index, NCI or ICI, with total scores vary according to the shapes of the total-score distributions. The findings are that if the total-score distribution has a negative skewness, then the correlation is positive; if the distribution is positively skewed, then a correlation results (Harnisch & Linn, 1981b; Tatsuoka & Tatsuoka, 1982b). Since the ECIs are natural extensions of the caution index, we can safely impute the same behaviors to these discrete personal indices as well. ECIs provide inflated values at both the extremely high and low total scores. With the standardized ECIs, the bias of the values at the extreme scores is corrected, and moreover the responses from different levels of θ can

be compared safely.

It would be ideal if the theoretical distribution of the standardized extended caution indices could be derived algebraically, but goodness-of-fit tests of ζ_1 and ζ_2 with normal distributions provide satisfactory evidence that they may follow approximately normal distributions (Tatsuoka & Tatsuoka, 1982b).

Regarding the detection rates of unusual response patterns, Rudner (1982) found that the personal indices (caution index, modified caution index, personal biserial index and appropriateness measures, person-fit statistics based on item response theory) detected only about 60% fo the unusual patterns within his Monte Carlo dataset. Tatsuoka and Tatsuoka (1982b) found that the detection rates of the response patterns resulting from application of erroneous rules by these personal indices including ζ_1 and ζ_2 are unexpectedly low, about 67%. By examining the frequencies of these rules with seven different datasets, it became clear that "popular rules" are difficult to spot by the use of personal indices. A popular rule is one used by a substantial number of students in the reference group from which the item parameters of logistic models are estimated. This implies that caution must be taken when applying the indices to practical situations to spot atypical response patterns.

However, the proptery of ζ_2 shown in Figure 4 enables us to find a student's misconception by first mapping his/her responses into the space spanned by θ and ζ_2 and th examining the region in which the corresponding point falls. If the student's responses are atypical, then his/her corresponding point will lie in the upper part of the space. If the responses are typical the point will fall closer to the horizontal axis, θ. As a result, it will be possible to distinguish popular misconceptions from unusual ones.

4. Tatsuoka,K. K. (1984b). A latent trait model for interpreting misconceptions in procedural domains. In D.J. Weiss (Ed.), the Proceedings of the 1982 Item Response Theory and Computerized Adaptive Testing Conference. Minneapolis, MN., 322-339.

Abstract

This study shows that item response theory is very useful in integrating cognitive processes into the theory and practice of educational measurements. For example, erroneous rules of operation in signed-number subtraction problems, resulting from various misconceptions, are represented as points in a geometric space by utilizing item response theory. We named this space "rule space." This approach seems promising in dealing with the variability of errors and assessing the state of knowledge appearing as an erroneous rule of operation.

5. Tatsuoka, K. K. (1985a). A probabilistic model for diagnosing misconceptions in the pattern classification approach. Journal of Educational Statistics, 12(10), 55-73.

Abstract

This paper introduces a probabilistic approach to the classification and diagnosis of erroneous rules of operation that result from

misconceptions ("bugs") in a procedural domain of arithmetic. The model is different from the usual deterministic strategies common in the field of artificial intelligence because variability of response errors is explicitly treated through item response theory. As a concrete example, we analyze a dataset that reflects the use of erroneous rules of operation in problems of signed-number subtraction. The same approach, however, is applicable to the classification of several different groups of response patterns caused by a variety of different underlying misconceptions, different backgrounds of knowledge, or treatment.

Discussion

This study introduced a new model that is useful in handling the response patterns of a given test together with various item response curves theoretically estimated from observed data. The model starts with mapping all response patterns into a Cartesian product space, $\{(\theta, f(x))\}$ (called rule space) using the continuous linear function, $f(x) = (x - P(\theta), P(\theta) - T(\theta))$. Quite a few response patterns are yielded by a variety of misconceptions that students possess. These response patterns associated with some erroneous rules of operation are also mapped in the rule space along with observed response patterns. Thus, examining the closeness of a student's position in the rule space to the positions representing the erroneous rules enables us to diagnose his or her misconception that resulted in the erroneous rule.

Actual responses, however, are often affected by "slips," or random errors produced by uncontrollable factors. When most deterministic models cannot handle random errors, the model utilizing item response theory enables us to take into account the variability of errors in responses to the items when diagnosing error types.

The pattern classification approach in the rule space is useful in avoiding some pitfalls from which the personal indices suffer (Tatsuoka & Tatsuoka, 1982, 1983), such as the problems of setting cut-offs to determine atypicality of response patterns, dependency on a baseline ordering of items, and ineffectiveness of detecting aberrations occuring in the middle positions of the order of the items.

It is crucial, however, to investigate the statistical properties of various clusters around each bug to utilize the powerful techniques developed in the area of statistical pattern classification and recognition.

6. Tatsuoka, K. K., & Tatsuoka, M. M. (1985). Bug distribution and pattern classification (Technical Report 85-3-ONR). Urbana, IL: University of Illinois, Computer-based Education Research Laboratory.

Abstract

A model (called rule space) which permits measuring cognitive skill acquisition, diagnosing cognitive errors, detecting the weaknesses and strengths of knowledge possessed by individuals was introduced earlier. This study further discusses the theoretical foundation of the model by introducing "Bug Distribution" and hypothesis testing (Bayes' decision rules for minimum errors) for classifying an individual into his/her most plausible latent state of knowledge. The model is illustrated with the domain of fraction arithmetic and compared with the results obtained from a conventional Artificial Intelligence approach.

This report was published in Psychometrika under the same title in 1987. The discussion of the article in Psychometrika is below:

7. Tatsuoka, K. K., & Tatsuoka, M. M. (1987). Bug distribution and statistical pattern classification. Psychometrika, 52(2),193-206.

Discussion

A new probabilistic model that is capable of measuring cognitive-skill acquisition, and of diagnosing erroneous rules of operation in a procedural domain was introduced by Tatsuoka and her associates (Tatsuoka & Tatsuoka, 1982). The model, called rule space, involves two important components: (a) determination of a set of rules to be diagnosed, or in other words, conditional density functions representing clusters around the rules, and (b) establishment of decision rules for classifying observed response pattern into one of the clusters around the rules and computing error probabilities. If each cluster around a rule can be described by a bivariate normal distribution of θ and ζ, the application of the techniques available in the theory of statistical classification and pattern recognition is fairly straightforward. With regard to the first component, a list of rules is supplied independently from parameter estimation of the item response theory models. Diagnoses of students' responses to the items are performed by classifying them into one of the bug distributions if possible, and if not possible are left to future investigation for searching the cause of misclassification. Determination of the list of the rules will be discussed in a future paper.

This study shows that, provided the slips can be assumed to occur independently across items, the cluster around the rules consisting of the response patterns resulting from one, two,..., several slips away from perfect application of the rule is well described by a compound binomial distribution with centroid (θ_R, ζ_R) and variance $\Sigma_{j=1}^{n} p_j q_j$, where $p_j (j = 1,...,n$ is the probability of having a slip from rule R for item j. The values of p_j and q_j are the logistic probabilities $P_j(\theta_R)$ and $Q_j(\theta_R), j = 1,...,n$, estimated from the dataset. Appropriateness of using the slip probabilities associated with each erroneous rule by the logistic function is left as a future topic of investigation, although the fit with the data seems to be good.

The determination of a set of ellipses representing clusters around the rules can be automatic once all the erroneous rules have been discovered. Many researchers in cognitive science and artificial intelligence have started constructing error diagnostic systems in various damains in this decade. Expert teachers usually know their students' errors, as well as the weaknesses and strengths of each child's knowledge structure. Since the model does not require large-scale computations such as strategies commonly used in the area of artificial intelligence do, the rule-space model is helpful in more general areas of research and teaching. Those who have microcomputers for testing their hypotheses are enabled to validate their data with probabilistically sound information, and to evaluate their teaching methods and materials. Moreover, the model can be "intelligent" in the sense that the researcher can improve and modify the information for the cluster ellipses as they

get more new students whose performance they can study.

The set of ellipses can represent many things besides erroneous rules. They can represent specific contents of some domain, usage errors in the language arts, or processes required in algebra. However, further research is necessary to develop methods for determining the set of ellipses other than relying on an expert teacher. The method must be efficient and compatible with recent theories of human cognition and learning.

8. Tatsuoka, K. K. (1985b). Diagnosing cognitive errors: Statistical pattern classification and recognition approach (Technical Report 85-1-ONR). Urbana, IL: University of Illinios, Computer-based Education Research Laboratory.

Abstract

This paper introduces a probabilistic model that is capable of diagnosing and classifying cognitive errors in a general problem-solving domain. The model is different from the usual deterministic strategies common in the area of artificial intelligence because the item response theory is utilized for handling the variability of response errors. As for illustrating the model, the dataset obtained from a 38-item fraction addition test is used, and the students' responses are classified into 34 groups of misconceptions. These groups are predetermined by the result of an error analysis previously done, and validated with the error diagnostic program written by a typical formal logic approach.

The published version of Technical Report 85-1-ONR has appeared in Behaviormetrika which is a journal of quantitative psychology, and technology in Japan. The Discussion section of this Behaviormetrika paper is given below.

9. Tatsuoka, K. K. (1986). Diagnosing cognitive errors: Statistical pattern classification and recognition approach. Behaviormetrika, 19, 73-86.

Discussion

A probabilistic model that is capable of diagnosing and classifying cognitive errors introduced in this paper. The model is constructed in the context of item response theory, and as a result, dealing with the variability of response errors becomes easier. Since all response patterns are mapped into a two-dimensional vector space spanned by MLE θ and ζ_2, which is defined as the rule space, this approach can be used for evaluating different problem-solving domain. The point is that the two different teaching methods will produce a significantly different number of students committing different types of bugs (Tatsuoka, Eddins & Sharabash, 1985), and hence they produce different clusters in the space. By examining these clusters closely, it is possible to evaluate the teaching methods and redesign them for more efficient instruction.

An important characteristic of the model different from other psychometric techniques such as cluster analysis, multidimensional scaling, and factor analysis is that it is possible to control the classification of performances on the test by selecting a meaningful set of ellipses. These ellipses represent erroneous rules in a procedural domain, or sources of misconceptions producing a variety or erroneous

rules of operation, or different achievement levels. Since users of the model control the input of ellipses according to their need or interest, they can avoid the tricky question about interpretation of factors or clusters. The quantity expressed by the ζ-axis represents the atypicality of a response pattern with respect to a given sample. Its distance from the θ-axis (MLE θ representing state knowledge) provides information concerning how likely it is that a particular bug will appear in the sample. This probabilistic interpretation is a unique feature of the rule space model which enhances the already useful prescriptive information about test performances with likelihood of such an incident.

Since the model is generalized version of mastery testing or criterion testing, it is applicable to the computer-aided instruction on micro-computers as an integrated part of training programs. Diagnosing error types instead of using the total score will be quite useful in training students. Moreover, the model can be used as an efficient placement tool for administrators of training programs. Mayberry, et al. (1985) demonstrated that the model places qualified students in various levels of courses in mathematics more efficiently, minimizes the number of dropouts and maximizes learning of a new topic for A- or B- students.

However, the new approach requires a couple of technical problems to be solved. Better estimation of ellipses will enhance the analysis results. A better parameter estimation technique will also provide efficient prescriptive information for each student. An accurate decision of classfying each point to one of the ellipses will be very important. These techniques are wide open for further research.

10. Tatsuoka, K. K. (1986). A cognitive error diagnostic adaptive testing system. Proceedings of the 28th International ADCIS Conference, 338-342.

Abstract

This paper introduces a new probabilistic approach for diagnosing cognitive errors by selecting an optimal set of items tailored to an individual examinee. The method is based on Item Response Theory and developed to a general theory so that it is applicable to any domain of interest. However, it will involve a painstaking task analysis.

11. Tatsuoka, K. K. (in press). Toward an integration of item-response theory and cognitive error diagnoses. In N. Fredericksen, R. L. Glaser, A. M. Leshold, & M. J. Shafto (Eds.), Diagnostic Monitoring of Skill and Knowledge Acquisition. Hillsdale, NJ: Erlbaum.

Abstract

This study introduces an outline of the cognitive error diagnostic methodology (rule space model) developed over the past several years by Tatsuoka and her associates. Objectives of the theory of cognitive error diagnosis are discussed first, and then pros and cons of latent trait and latent class models are discussed in terms of four dimensions of the goals to be attained in modern educational measurement theory.

This paper is intended to clarify the philosophy behind the rule space model and tries to relate it to one of the fundamental issues of educational measurement--what really determines the item response curves, a question which is somewhat similar to Carroll's quest for "what ability is." A new technique that is an application of Graph Theory is

introduced to determine a list of bugs--sources of misconceptions, locations of incomplete knowledge structure, or various combinations of task attributes relevant to the solution of problems--to be diagnosed.

The rules listed in the bug information bank are translated into multivariate normal distributions (bug distributions) in the rule space and their behaviors are discussed in the context of statisticaal properties of the bug distributions. Finally, it is shown that the rule space is useful in diagnosing different strategies adopted by students for solving the same problems, together with diagnosing individual cognitive errors. Diagnoses are carried out by using Bayes' decision rule for minimum error after computing the Mahalanobis distances of the student's point to the centroids of the ellipses representing the bug distributions

Discussion

This paper discussed some important issues for theories in educational measurement and testing. Recent findings in modern learning theory have raised tough challenges to psychometricians. Glaser categorized these challenges into four main objectives that should be taken into account by new achievement measurement. These objectives are descriptive, dynamic, structure-oriented, and orchestrate several component tasks.

The pros and cons of two representative approaches of probabilistic modeling commonly used in psychometrics were discussed. Their common, basic principle is that an individual's proficiency level is explained, to a substantial degree, by defining certain human characteristics called traits. These traits are invisible. The only observable outcomes are the students' responses to the test items.

Modern physics and advances in theory and practice of electricity and electronics share this problem with us. We have to infer the outcomes of invisible traits (if they exist), unobservable cognitive processes, individuals knowledge structures and theory changes from observable responses to test items. Modern physicists have discovered neutrons, electrons and other elementary particles by modeling observable physical phenomena as logical consequences of the postulated properties of the unobservable elementary particles..

It would seem that a new addition to the current theory and techniques of psychometrics is needed to incorporate the new challenges raised by modern learning theory. This addition must overcome an ideological barrier as well as a technological one as stated by Linn (1985). At the same time, the new addition requires an expansion of our common sense to a more abstract level, and a re-examination of the basic principles of test theories. However, the time is ripe, and finally research that manifests a new trend has begun to appear in several journals. This includes, for example, renewed concepts of construct validity proposed by Messick, Angoff and Cronbach, and a new technology, called rule space, introduced by Tatsuoka and her associates.

This study introduced a part of the rule space model. The model consists of four major components: (1) Item construction and preparation of the bug library. The bug library consists of response patterns resulting from various sources of misconceptions, sources of incomplete knowledge and erroneous rules of operation: (2) Estimation of parameters, including item parameters of IRT models and bug distributions; (3) Execution of decision rules; (4) Evaluation of the information in the bug

library and update of the contents.

Each component requires a substantial amount of time to discuss in detail. The first component was discussed mainly by introducing a method for constructing an item-tree. Making a tree is an application of deterministic relational databases. Each item tree reflects a unique process structure underlying the problem-solving activities. The values of conditional probabilities computed on a specific directed path will help to identify the attribute that causes difficulties in doing the test.

After locating a source of error types, or combinations of several attributes that require special attention in teaching, remediation, or designing instructions, the item-tree program converts them into a list of response patterns. Since the attributes can be production rules, the item tree can be a descriptive representation of a production system.

One of the main differences between the traditional modeling approach and the rule space approach lies in the ways they utilize the information obtained from detailed task analysis. The former approach defines a set of new variables and formulates them as parameters in a probability function, or formulates them into probabilistic relationships among probability functions. The rule space approach utilizes algebraic relationships among item response functions for expressing the information obtained from the task analysis. Rules are associated with bug distributions and represented as points in the rule space which is a Cartesian product of two quantities, θ and ζ. Each rule forms the center of an enveloping ellipse with the density of points getting rarer as we depart farther from the center in any direction. Further, the major and minor axis of these ellipses are asymptotically orthogonal (Tatsuoka, 1985). An observed response, on the other hand, will be classified into one of the ellipses if possible. Statistical pattern recognition techniques are applied to classify the observed point. By examining the probability of errors, the student's performance on the test will be diagnosed with an interpretable prescription.

Since the meaning of θ is taken as denoting the levels of proficiencies and not as latent traits or constructs which govern obtaining certain levels of performances on the tests, the change of scores resulting from hypothesis-testing activities of tests are explained smoothly without any philosophical difficulties. The model has treated θ as a quantitative variable and contributed an important role to the contents of the bug library.

II. Applications: Learning, Cognitions, and Measurement Issues

1. Tatsuoka, K. K., & Tatsuoka, M. M. (1983). Spotting erroneous rules of operation by the individual consistency index. Journal of Educational Measurement, 20(3), 221-230.

Abstract

This study introduces the individual consistency index (ICI), which measures the extent to which patterns of responses to parallel sets of items remain consistent over time. It can therefore be used as an error diagnostic tool to detect aberrant response patterns resulting from the consistent application of erroneous rules of operation.

The function of ICI, which is different from other personal indices, is discussed and illustrated with data from achievement tests of a

problem-solving type. However, a drawback of ICI is that it requires repeated measures in a test. As a result, ICI is not applicable to most standardized tests and its use is limited to tests that are deliberately constructed with parallel parts.

Discussion

Although the content area in this study is a specific domain, signed-number arithmetic, ICI should be useful in measuring the stability of a student's algorithm throughout a test in any domain, and should also be useful in spotting the student's systematic misconceptions. The unique feature of ICI is that its values are individually oriented and free from the group dependence that can affect indices such as NCI and C_i. It reflects a student's own procedural skill. Since the development of systems such as SIGNBUG is very expensive and time-consuming, spotting any response patterns having high ICI values and low total scores will, to a certain extent, serve as a substitute for expensive error diagnostic systems.

Although ICI is useful in detecting aberrant response patterns resulting from the use of wrong algorithms, it requires repeated measures in a test. Therefore, the index is not applicable to many commercial achievement tests or to criterion referenced tests designed to measure the outcome of treatments in a wide range of content areas. In these cases, the judicious use of C_i may be indicated. However, when tests are aimed at assessing the progress of students' learning activities and used as an integral part of instruction, the information obtained from ICI values will be useful for tapping the students' degree of understanding of the subject matter. This line of research will be much more necessary as computer technology penetrates into educational usage as Glaser and Nitko advocate in their recent papers (Glaser, 1981; Nitko, 1980).

2. Harnisch, D. L., & Tatsuoka, K. K. (1983). A comparison of appropriateness indices based on item response theory. In R. K. Hambleton (Ed), Applications of Item Response Theory, 104-122

Abstract

The focus of this study is on appropriateness indices based on item response theory (IRT). Fourteen appropriateness indices are introduced and reviewed. The interrelationship among the appropriateness indices reveals a relatively high level of association with over 50 of the correlations being above an absolute value of .7. The group of standardized extended caution indices was related the least with the total test score for data from the National Assessment of Educational Progress for 13-year-old students in mathematics. Tests for the curvilinearity of the appropriateness indices were conducted and indicated that the group of standardized extended caution indices was least resistant to change when a square and cubic term was added to the model. Applications of appropriateness indices to educational measurement are noted.

3. Tatsuoka, K. K. (1984). Changes in error types over learning stages. Journal of Educational Psychology, 76(1), 120-129.

Abstract

Studies of procedural errors in arithmetic have shown that a variety of misconceptions of "bugs" may yield some 200 erroneous rules of operation. This study attempts to classify the bugs discovered in signed-number addition problems into several categories to facilitate the investigation of changes in these errors at different times and the designing of remedial instructions. The classification was first done by examining which of the procedural steps were correctly completed by each erroneous rule. Second, an Index was defined that measures how early or how late in the procedural tree (or network) an erroneous rule causes departure from the correct steps; an early derailment obviously has more serious consequences than one at a later stage. Erroneous rules formed because of a misconception or lack of understanding at one or two earlier stages of the procedural network formed a single group and showed properties different from all other rules.

Discussion

The 27 erroneous rules of signed-number addition problems were classified into two groups, serious error (Group B) and non-serious error (Group A) types, and the changes in their incidence rate over time was investigated. The rules in Group B are due to the lack of understanding or misconceptions at one or more deeper levels of the procedural steps. As a result, the rules are ill-composed and applied pretty consistently throughout the test (Tatsuoka, 1981). It is interesting to note that the percentage of Type B rules over the six tests is monotonically decreasing regardless of the treatment. It seems that the two treatments--an hour of PLATO lessons versus intensive, 3-week, regular classroom instruction--do not make much difference in the incidence of these ill-composed rules. Perhaps mental maturity, or the gaining of other experiences by the students is the main cause of the declining frequency of serious errors over the time period. On the other hand, a specific remedial instruction designed for correcting each ill-composed rule in Group B may produce positive results. Further investigation is needed to explore efficient methods to remedy these serious misconceptions.

Finally, we investigated the transitional behavior of rules of the six groups--right rule, other, incomplete rules of Groups A and B, and complete rule of Group A and B--from one test to the next. The results showed that all kinds of transitions occurred haphazardly. In order to discern anything systematic in the transitional behavior, future studies will need to observe far more subjects than we have done to date.

4. Tatsuoka, K. K., & Eddins, J. M. (1985). Computer analysis of students' procedural "bugs" in an arithmetic domain. Journal of Computer-based Instruction, 12(2), 34-38.

Abstract

This study investigated the transition of various erroneous algorithms over different stages of learning in signed-number subtraction problems. The results of analysis by an error-diagnostic computer program revealed that even if a substantial amount of practice and drill was given to students, ambiguity existing in the instructional material affected test performance. It is important that students understand the notation used in the test and that a smooth transition be made from the manipulative tools used in teaching concepts and procedural skills into rules of operation (algorithms).

Summary

In summary, this study demonstrated that error analysis can provide information useful for correcting weakness in instruction and for prescribing remediation for specific misconceptions. Use of instructional techniques such as number lines is important, and many educators have invested effort in finding ways to express structural understanding of numbers. However, the parallel protocol studies conducted in this experiment indicated that most students did not use the number-line method to answer test items even during the several exercises given to them on the PLATO system before they took Test 4.

They struggled to hypothesize and invent their own algorithms as soon as the manipulation of the pointer on the line was introduced. When asked by the interviewer, "Why don't you use the number line?", one student answered, "Because it takes too much time." He invented a wrong algorithm for - L - (- S) and -S - (- L) item types. His rule was, "The minus operation sign has the power of changing all of the minus signs to plus signs." Thus, his rule changed - L - (- S) to + L + (+ S), and - S - (- L) to + S + (+ L). His error disappeared by the time he took Test 5.

In fact, most errors disappeared by Test 5, as can be seen in the line for subset U in Figure 2, and in all of the lines in Figure 4; however, those errors caused by ambiguous or incomplete instruction tended to persist longer. Thus, a student may acquire an understanding of numbers through the invention of his or her own rules, but the process can be speeded up if reinforced by an analysis of how those rules work.

5. Birenbaum, M. (1986). Effect of dissimulation motivation and anxiety on response pattern appropriateness measures. Applied Psychological Measurement, 10(2), 167-174.

Abstract

This study examined the effect of anxiety and dissimulation motivation of job applicants on their performance on an ability test. Two aspects of performance were considered: the total score and the appropriateness score. Four IRT-based appropriateness indices for detecting aberrant response patterns were employed in this study. The results indicate a negative effect of dissimulation motivation on the performance of low anxiety scorers, with respect to both the total score and the appropriateness score, with a greater effect on the latter. This effect was evidenced by an erratic or aberrant response pattern on the ability test; that is, missing relatively easy items while answering more difficult ones correctly. The results are discussed in light of the diverse interpretations concerning the meaning of Lie scales.

Discussion

The negative correlation between Lie and Pt scores in the present study indicates, according to the criterion suggested by Michaelis and Eysenck (1971), the existence of dissimulation-motivating conditions under which the examinees responded to the questionnaires. Since anxiety is particularly susceptible to dissimulation, it is reasonable to anticipate that, in addition to genuine low-anxiety examinees, the group of low anxiety scorers includes some high-anxiety examinees who tried to impress or deceive their assessors by reporting a low anxiety level in

order to get the jobs for which they applied. This interpretation, which follows the most common explanation of the meaning of the Lie score, seems to have gained substantial support from the results of the present study. There was a significant difference in ability performance between low and high Lie scorers in the low-anxiety group. Moreover, there was no significant difference in ability performance between the latter group and the high-anxiety groups.

This interpretation leads to the conclusion that an examinee may be able to falsify his or her responses on an anxiety test, but not the effect of anxiety on his or her cognitive functioning. This effect is expressed in an erratic or aberrant response pattern on the ability test (i.e., missing relatively easy items while answering more difficult ones ocrrectly), as can be seen from the comparison of item difficulties in the two groups. The attentional interpretation of anxiety provided by the cognitive approach helps to clarify this effect, which is especially noticeable in evaluating situations. According to Green (1976, 1980) and Sarason (1984), it is the worry over evaluation that leads to task-irrelevant cognitions that interfere with attention to the range of cues in the situation.

Supporters of the competing explanation concerning the meaning of Lie scales may argue that it is the compulsiveness aspect, characteristic of a high-conforming person, that is responsible for the aberrant response patterns in the low anxiety-high Lie group. According to Cattell et al. (1970), one of the main contributors to the second order factor of conformity is compulsiveness (Q3), which is described as "the extent to which the person has crystallized for himself a clear, consistent, admired pattern of socially approved behavior, to which he makes definite efforts to conform" (p.107). Gordon (1985) proposed a theory which predicts that obsessive-compulsive disorder is characterized by "hyperattention and that under mild stress these subjects would therefore show disruption of attention, particularly in controlled information processing" (p. 105). When stress is increased this would cause, according to Gordon, "greater deterioration in performance in these subjects than in other groups, seen in reduced accuracy and slower reaction" (p. 105). The fact that the high anxiety-high Lie group scored the highest on the inappropriateness measures, and the lowest on the total and ability scores, seems to support this interpretation.

Taken together, it seems that both anxiety and compulsiveness contribute to the inappropriateness of the responses. Anxiety is particularly dominant in the high-anxiety-low Lie group, compulsiveness in the low anxiety-high Lie group, and both in the high anxiety-high Lie group. The fact that anxiety and compulsiveness affect performance, due to attentional disturbances which result in missing easy items on the ability test, should be taken into consideration when assessing the ability of examinees exhibiting those personality characteristics. The implications for test practitioners are quite obvious. In assessing an examinee's ability in an evaluating situation, the appropriateness of the response pattern must be considered as well as the total test score.

The four indices used in the present study for measuring response pattern appropriateness yielded similar results. Their high intercorrelations confirm previous results (Harnisch & Tatsuoka, 1983; Rudner, 1983; Birenbaum, 1985). They can, therefore, be recommended as altermate measures of response pattern appropriateness.

6. Shaw, D. S. (1986). Cognitive error diagnosis and prescription for fraction arithmetic. Proceedings of the 28th International ADCIS Conference, 324-331.

Abstract

Much of the research in computerized instructional design has focused upon aspects of the instruction itself. While such studies are important to improve the effectiveness of computer-based education, attention needs to be centered upon characteristics of the learner. In this study, it was found that brief interactive dialogue could increase scores of students who were diagnosed, using the rule space approach, as having a particular misconception. The instruction was not as worthwhile for students whose needs were different.

Discussion

The mathematical model, Rule-space, makes it possible to generalize the adaptive testing procedure as well as map the outcome of instruction. The development of appropriate test items is basic for two reasons. First, the itsm parameters are a requirement for the determination of the true score (the "x" axis in the rule-space concept), and the items serve as a tool for making observations and structuring the item tree.

For example, present work in teaching the use of automated graphics programs of architectural design has the same kind of item requirements even though the subject domain is different. Questions concerning the processes of representation of space and line in a computer must be developed in order to measure the user's conceptual level, in much the same way that questions were developed in fraction arithmetic. Studies of wrong approaches or inefficient methods of work can lead to a skills hierarchy or task structure. Preliminary protocol studies reveal differences in mental models of specific classes of user. The use of "shortcuts" seems to be characteristic of "experts" or those who really understand, as observations of borrowing procedures often revealed. All subject domains have similar attributes which must be discovered.

Of particular value in the rule-space concept is the mapping of changes in order to describe and evaluate the effects of instruction. Without graphic representation, it would be difficult to draw conclusions concerning the kinds of conceptual shifts occurring with different instructional strategy, which is essential to remediation.

7. Birenbaum, M., & Tatsuoka, K. K. (1987). Effects of "on-line" test feedback, on the seriousness of subsequent errors. Journal of Educational Measurement, 24(2), 145-155

Abstract

The present study examined the effect of three modes of feedback on the seriousness of error types committed on a post-test. The measure of seriousness of error types, or wrong rules indicates to what extent the wrong rules deviate from the right rule. An on-line free-response six-item pre-test covering addition of signed number was administered to 263 8th graders. Upon completion of the pre-test the students were randomly assigned to receive one of the following kinds of feedback to responses to a second six-item test: 1) information about whether or not the response was correct, 2) the correct answer, or 3) the correct rule for solving the problem. The effect of the feedback mode on the

seriousness of errors committed on a third six-item test, the post-test, was found to be differential and dependent upon the seriousness of errors committed on the pre-test.

Discussion

The effect of the three feedback modes on the seriousness of errors committed on the post-test was found to be differential and dependent upon the seriousness of the errors committed on the pre-test. Feedback seems to have little or no affect on pre-test serious errors. They remain unchanged for the most part. The same trend has been noticed for correct responses on the pre-test. Nonserious errors, on the other hand, tend to be more responsive to certain modes of feedback than others. The more informative the feedback is with regard to the correct answer and correct rule, the more likely are the post-test results to be correct. The highest rate of transistion from nonserious to correct responses occured in the third (CR) treatment group, which received the most informative feedback mode; i.e., the correct rule was provided to all the students in that group in addition to a diagnosis of the erroneous rule where is was consistently used across all the test items, in which case it could be identified by the error diagnostic program. (Only 28% of the students committing errors on the pre-test received that type of feedback). The second highest rate in transition from nonserious to correct answers occurred in the second (CA) treatment group which received the correct answer. The lowest rate occurred in the (OK) group whose feedback mode provided only confirmation for correct responses.

These results are in line with existing knowledge concerning the relationship between feedback process and learning. The information processing approach to the study of feedback supports our finding that the feedback turned out to be ineffective in te case of serious errors. As stated by Kulhavy (1977) "the feedback acts to inform the student concerning the accuracy of his response relative to knowledge he already possesses about the content. Obviously, if the material studied is unfamiliar or abstruse, seeing feedback should have little effect on criterion performance, since there is no way to relate the new information to what is already known" (p. 220).

Our findings also support Kulhavy's claim that feedback for incorrect responses is more important than for correct ones -- an argument that disputes the operant psychologists' view of feedback as a reinforcement event (e.g., Skinner, 1986).

The finding that the more informative the feedback mode, the better the performance, is in agreement with previous findings (e.g. Gilman, 1969; Holland, 1965). However, our findings indicate an essential distinction between the type of errors that are more likely and those that are less likely in their facility to be affected by the appropriate feedback mode.

The most informative feedback mode in our study was meant to produce an adaptive feedback. However, the fact that we used a deterministic diagnostic system impaired the achievement of this goal, since our system was only capable of identifying the rule of operation systematically applied and was able to identify only 35% of the incorrect responses. The fact that the remaining 65% of the responses were inconsistent, and therefore remained unidentified, does not mean that they were all careless or random errors. Studies of thinking suggest that learning results from a process of hypothesis testing (Bruner, Goodnow & Austin,

1956; Levine, 1966; Mayer, 1983); thus, one can expect students to change their strategies during the test. In fact, the "repair" theory by Brown and VanLehn (1980) predicts the change in students' strategy when confronted with an "impass" (VanLehn, 1982). A probabilistic approach to error-analysis (Tatsuoka, 1984; 1985) there is therefore more appropriate.

Such an approach, unlike a deterministic one, enables one to diagnose an underlying misconception in the presence of "careless" errors or slips, and thereby increases the identification rate of students' "bugs". It seems that application of probabilistic models of this kind for diagnosis can provide the necessary means for designing a more efficient feedback mode which is better tailored to the students' needs.

8. Birenbaum, M., & Tatsuoka, K. K. (1983). The effect of a scoring system based on the algorithm underlying the students' response patterns on the dimensionality of achievement test data of the problem solving type. Journal of Educational Measurement, 20(1), 17-26.

Abstract

This study focuses on issues related to the dimensionality of achievement test data of the problem-solving type. Two methods for scoring test data on a free-response signed-number test were used. One scoring method was based on results of an error analysis and credited only correct answers presumably derived by the correct algorithm. The second scoring method was the conventional one which credits all right answers. The outcomes of the two scoring methods were compared in terms of reliability and dimensionality.

The results indicate that the conventional method was inferior to the other one in both aspects: it resulted in negative correlations among items and a lower reliability coefficient. The scoring method based on the underlying algorithm resulted in a higher reliability coefficient and a smaller and more "crystallized" dimensionality than the conventional scoring method.

Discussion

The idea that students tend to modify the algorithm taught in class is already a well-established one among cognitive psychologists. The fact that some of these modifications result in incorrect algorithms should concern psychometricians as well, especially in the cases where wrong algorithms happen to yield correct answers, as was often the case in our study.

The results of our study indicate that in achievement data of the problem-solving type, where a specific subject matter area is being tested, the factorial structure of the data is highly affected by the existence of different algorithms underlying the student response patterns. These systematic sources of variation cause an increment in the dimensionality of the data. The fact that students can sometimes get right answers by following a wrong rule is reflected in the psychometric properties of the test. Thus, when the conventional scoring system is used, it results in negative correlations among some items, and increased dimensionality. When a scoring system that takes into consideration the process rather than relying solely on the outcome of the cognitive process is used, significant gains in the psychometric properties of the

test can be obtained. The structure of the test crystallizes as can be seen in Figure 2. Thus, while the amount of modification made was relatively small (less than 6% of the scores having been changed), its effect on the psychometric properties of the test was quite noticeable. Moreover, the variation in the amount of modification made in the different tasks is not highly related to the amount of changes that occur in the task-total correlations as result of the modification. Therefore, one cannot substitute the search for algorithms with the conventional discrimination index of the classical test theory.

The results of this study may have some important implications for psychometric work on achievement tests of the problem solving type. "Latent trait" has become a very popular concept in the last few years among psychometricians. Sophisticated mathematical models are being developed for estimating ability from responses to test items (Lord & Novick, 1968). Adaptive testing, in which the items are meant to be tailored to the examinee's level of ability, is constructed using the estimates derived from those models. It seems that transferring this approach to achievement tests needs modification due to the different nature of achievement testing from ability testing. Be the purpose of the achievement test diagnosis or be it for estimating achievement level in selection/classification programs, achievement tests are always measuring an outcome of a treatment (i.e., instruction). That treatment was meant to provide the student with an algorithm to be used in solving problems in a specific subject matter area. The student's responses on the test items reflects his/her modification of that algorithm. The aim of the testing should therefore be to identify that latent cognitive process in order to quantify the responses in a more accurate and at the same time more meaningful manner. Adaptive tests in this context should aim at "debugging" the student's misconceptions, thus enabling a better understanding of his/her algorithm or strategy in solving the problem.

9. Birenbaum, M., & Shaw, D. S. (1985). Task specification chart: a key to a better understanding of test results. Journal of Educational Measurement, 22(3), 219-230.

Abstract

A task specification chart (TSC) that integrates the content facets and the procedural steps of a specified task is suggested as a tool for designing a test and for interpreting its results. An instructional unit for adding and subtracting fractions was used to demonstrate the design and application of the TSC. To evaluate its efficacy, a test based on the TSC was administered to two independent samples. Accounts of the variance of item difficulty indices and errors in students' responses were used as the criteria for this evaluation. The results indicated that a very high percentage of variance in item difficulty indices was accounted for by item characteristics representing different task components. The typology of errors constructed on the basis of the TSC proved to be an efficient tool for identifying erroneous rules of operation underlying students' response patterns on the test.

Discussion

The gap between testing and instruction has been one of the main concerns of educators and psychologists for the last decade. The study reported here focuses on the effect of a task specification chart and its

implications for testing and instruction. The results indicated that item characteristics representing different task components can account for most of the variance in item difficulties on tests. Moreover, a stable pattern of relationships between the different item characteristics and the variance in item difficulty was indicated across samples which differed significantly in total test scores.

The fact that the empirical results confirmed the relationships among the components specified in the chart has important implications for testing as well as for instruction. Using a task specification chart, the teacher can design instruction to match the difficulty of the task components. The chart also provides a basis for describing students' levels of achievement in terms of componsent not mastered.

An attempt was also made in this study to diagnose students' errors. Burton (1981) defined diagnosis as determining the internalized set of incorrect instructions or rules that gives results equivalent to the students' results. Glaser (1981) stressed the importance of identifying performance irregularities; they inform the activities of instructional decision making and student guidance.

The typology of errors identified in this study should prove to be useful tool for diagnosis. The classification of errors according to the task components should enable the teacher to evaluate instruction and redesign it accordingly. The analysis of a student's response patterns in terms of the underlying rules of operation enables the teacher to prescribe individualized remediation to address the misconceptions reflected in that student's responses. This kind of information is above and beyond the kind provided by the traditional method of scoring a test, that is, for the total number of correct responses. As can be seen in Table 4 and elsewhere (Birenbaum & Tatsuoka, 1982; Tatsuoka & Tatsuoka, 1983), the same total score can represent completely different misconceptions. In view of this, the task specification chart promises to be an effective tool in the area of testing, both for test design and for diagnostic assessment of students' performance. Its use can also establish a reciprocal relationship between testing and instruction.

10. Birenbaum, M., & Tatsuoka, K. K. (1987). Open-ended versus multiple-choice response formats--it does make difference for diagnostic purposes. Applied Psychological Measurement, 11(4),

Abstract

The purpose of the present study was to examine the effect of the response format: open-ended (OE) versus multiple choice (MC) on the diagnosis of students' misconceptions in a procedural task. A test in fraction addition arithmetic was administered to 285 8th-graders, 148 of whom responded to the OE version of the test and 137 to the MC one. The two datasets were compared with respect to the underlying structure of the test, the number of different error-types and the diagnosed sources of misconception (bugs) reflected in the response patterns. The overall results indicated considerable differences between the two formats with more favorable results for the OE format.

Summary and Discussion

The purpose of the current study was to examine the effect of the item format (OE vs. MC) on students' responses to a procedural task test in fraction addition arithmetic. The overall results indicated

considerable differences between the two formats with more favorable results for the OE format. The underlying structure, as examined by Smallest Space Analysis, seemed clearer in the OE dataset, where the configuration of the items in the two dimensional space clearly indicated two clusters; one of items with like denominators and the other of items with unlike denominators. The item configuration for the MC dataset, on the other hand, seemed quite diffused with no distinct separation between the different item types.

The results of the error-analysis provided an even clearer distinction between the two response-formats. Although the two groups did not differ in the ability to solve fraction addition problems, the MC dataset included a significantly larger number of different error types than the OE dataset. This resulted in a less appropriate overall classification rate in the Rule Space. These results seem to indicate that students who have not mstered the task tend to be less consistent in applying their rules of operation for solving procedural tasks when faced with an MC format than with an OE one. It seems that the cognitive process involved in these two response formats (in procedural task) is quite different. According to Fisher and Lipson (1985) "Humans exhibit a fairly strong tendency to avoid extra mental effort, so as to minimize their information processing load" (p.65). While in the OE items subjects had to compute the answer "from scratch," in the MC test they could retrieve cues from the distractors, thus shortening the process, perhaps with more efforts directed towards judging the "correctness" of the answer given by the distractors rather than towards carrying out the entire "tedious" calculation. However, since the distractors in the MC test were carefully chosen to represent common errors rather than random incorrect answers, the task of selecting the right answer became more complicated and resulted in a greater variety of error types, a greater part of which were rationally uninterpretable.

The implications for diagnostic achievement testing in procedural tasks is quite obvious. MC tests, though considerably easier to score, may not provide the appropriate information for identifying students' misconceptions with respect to the given subject matter. The OE format seems more appropriate for this purpose.

11. Tatsuoka, K. K. (1987). Validation of congnitive sensitivity for item response curves. Journal of Educational Measurement, 24(3), 233-246.

Abstract

This study sought a scientific way to examine whether item response curves are influenced systematically by the cognitive processes underlying solution of the items in a procedural domain (addition of fractions). Starting from an expert teacher's logical task analysis and prediction of various erroneous rules and sources of misconceptions, an error diagnostic program was developed. This program was used to carry out an error analysis of test performance by three samples of students. After the cognitive structure of the subtasks was validated by a majority of the students, the items were characterized by their underlying subtask patterns. It was found that item response curves for items in the same categories were significantly more homogeneous than those in different categories. In other words, underlying cognitive subtasks appeared to systematically influence the slopes and difficulties of item response

curves.

Discussion

Cognitive processes underlying students' performances on tests influenced the IRT curves in a well-defined procedural domain, fraction arithmetic. The items requiring the same set of cognitive subtasks underlying the correct solutions had similar curves; items requiring different subtasks tended to produce dissimilar curves. Although this conclusion was based on the study of fraction addition problems, the same result has also been obtained from other well-defined domains such as fraction subtraction and signed-number subtraction problems.

The conclusion indicates that the use of IRT for developing probabilistic models for measuring cognitive processes and knowledge acquisition should be encouraged. There may exist several different ways for utilizing IRT, but whatever the methods there are, it is important to expand our view of IRT from the current level of using IRT models for selecting testees or predicting their performance to new levels of diagnosing cognitive skills or assessing knowledge acquisition.

Many researchers have been working on IRT in order to increase the accuracy in estimating ability levels, and item parameters; but the interpretability of estimated θ-values has never been investigated or, at least, the majority of measurement researchers have ignored it (Glaser, Lesgold, & Lajoie, 1986; Messick, 1984). Nevertheless, Tatsuoka and her associates have developed the probabilistic model called Rule Space for exploring the interpretability of data (Tatsuoka, 1985, 1986, in press, Tatsuoka & Linn, 1983; Tatsuoka & Tatsuoka, 1982, in press). The rule space utlizes a variety of different shapes of item response curves for diagnosing sources of misconceptions or erroneouss rules of operation in procedural domains. Their approach may be just one of several possible ways to attain educational goals mentioned earlier in this study (Glaser, 1985; Linn 1985). Since IRT curves are sensitive to underlying cognitive processes involved in answering the correct responses to items, further mathematical and statistical properties of IRT must be explored rigorously in the future.

III. Order Analytic Approach

1. Tatsuoka, M. M. (1986). Graph theory and its applications in educational research: A review and integration. Review of Educational Research, 56(3), 291-329.

Abstract

This paper presents a non-technical exposition of graph theory (more particularly, the theory of directed graphs or diagraphs), followed by a survey of the literature on applications of graph theory in research in education and related disciplines. The applications included order-theoretic studies of the dimensionality of datasets, the investigation of hierarchical structures in various domains, and cluster analysis. The number of papers applying graph theory was found to be relatively small except in sociology. Possible reasons for the paucity of applications in educational research are discussed, and the value and feasibility of achieving increased use of graph theory in this field are also pointed out.

Concluding Remarks: Future Prospects

We have seen that the literature on applications of graph theory in educational and psychological research is not very large. Also, the kinds of application are limited mainly to three types: extraction of unidimensional chains of test items, construction of hierarchies of test items or instructional material, and clustering of objects. Whatever applications are made, however, the utility and effectiveness of graph theory were very much in evidence. Briefly stated, the main advantage gained by utlizing graph theory (and the related concepts of order analysis) is that it decreases the amount of subjective judgement required in achieving the goal--be it constructing item chains, item hierarchies, or clusters of objects. To be sure, the subjective element is never completed eliminated (nor would that be desirable even if it could be done) as we have repeatedly seen above. The choice of cutting points, the level of tolerance for inconsistent data, the selection of criteria for similarity or proximity all involve subjective judgments that cannot be avoided, or that we prefer not to eliminate even if it were possible. The point is that once the basic subjective judgments are made (be they based on intuitionn, expert knowledge, or logical analysis), the theory does the rest. This is perhaps most evident in connection with the extraction (or construction) of hierarchies of test items, instructional materials, and so forth. Subjective, substantive judgments need only be made in conjunction with pairs of objects that stand in an immediate relation of is a prerequisite to, and so forth. Once this is done, we have the basis for constructing an adjacency matrix, which can be operated on in accordance with the theorems of graph theory to yield the structure of mediated relations.

Why, then, is graph theory not more extensively applied to eduactional and related research? It is not surprising that the extent of application there is not as large as in electrical engineering and computer science, for instance. These fields are, after all, eminently cut out to utilize graph theory with which they share such concepts as networks, trees, circuits, and flows. In fact, many of the recent developments in graph theory have owed their origin and stimulus to substantive problems in electrical switching theory, electronic circuitry, and computer and information science. But how do we account for the more abundant application in such fields as sociology compared to education and psychology? Here again, the greater amenability of such concepts as social structure, communicatin networks, power structure in social hierarchies, and (especially) sociometric diagrams to graph-theoretic treatment is no doubt at least partly responsible for the greater popularity of graph theory in that discipline. (Support for the notion that applications of graph theory are more widespread in sociology than in education and psychology can be gleaned by scanning recent issues of such journals as the Journal of Mathematical Sociology and Social Networks, as well as periodic publications such as Annual Review of Sociology and Sociological Methodology.)

One is led to speculate, however, that another reason for the relative paucity of educational and psychological applications of graph theory may lie in the fact that phenomena in these fields are more inherently statistical and stochastic than are sociological phenomena. This is true especially in the field of learning, for learning is intrinsically and necessarily a stochastic phenomenon. But the study of

stochastic graphs is a very recent development (see, e.g., Fienberg & Wasserman, 1981; Frank, 1981; Holland & Leinhardt, 1977, 1981). According to Frank (1981), the "marriage" between graph theory and statistics can take three forms: graph sampling, graph transitions, and graph processes. The last, which deals with "graphs or diagraphs that change in time in [a probabilisitc] way" (p. 127) is possibly the most relevant to educational and psychological research that concerns learning phenomena. The latter have long been associated with Markov processes, and graph processes can be regarded as a continuous-time Markov processes in which the states correspond to adjacency matrices whose elements change probabilistically with time.

Alternatively, an ordinary n-state Markov process (not a graph process) itself may be represented by an n-point, weighted, connected diagraph. Here the points correspond to the states (such as the degree of learning, forgetting, concept acquisition, etc) and the weights of the arcs correspond to the transition probabilities between adjacent points (i.e., states). The reason this application of graph theory was not widespread until fairly recently is that the whole topic of weighted graphs is a relatively new development (see, e.g., Deo, 1974, pp.424-439, and the references given there. A weighted graph is one in which the arcs are assigned numerical values: cost, intensity, and so on,are examples of weights other than transition probabilities that can be associated with arcs).

If the above speculation has any merit, then the relatively scant application of graph theory in educational and psychological research may be attributed at least partly to the recency of the development of the stochastic graphs and weighted graphs. Each in its own way seems most relevant to the type of phenomena dealt with in these disciplines: stochastic graphs because they describe relations that change probabilistically with time, and weighted graphs (in the special case when the weights are transition probabilities) because, although the graphs themselves are static, they offer alternative and potentially useful representations of Markov processes. Now that these branches of graph theory are on the rise, there seems to be little reason for educational researchers not to take more advantage of these tools in their work.

2. Wise, S. L., & Tatsuoka, M. M. (1986). Assessing the dimensionality of dichotomous data using modified order analysis. Educational and Psychological Measurement, 46, 25-301.

Abstract

Previous research has shown that order analysis and factor analysis typically yield different results in assessing the dimensionality of dichotomous data. This paper demonstrates that, through the inclusion of item proximity information, the dimensions identified by order analysis can be highly congruent with the results of factor analysis. A modified order analysis procedure is presented and compared to factor analysis using both simulated data of known dimensionality and real data.

IV. Site Curriculum Development

1. Tatsuoka, K. K., & Yamamoto, K. (1985). Application of component

scoring to a complicated cognitive domain (Technical Report 85-2-ONR). Urbana, IL: University of Illinois, Computer-based Education Research Laboratory.

Abstract

The recent development of cognitive psychology and science suggests that the lack of understanding of important structural relations between the entities in a problem-solving domain causes difficulties in learning. This study proposes a new scoring method by which the structural relations as well as processes used in subcomponents of the knowledge structure are taken into account when determining the scores. By so doing, a score of "1" derived by wrong reasons will be eliminated and patterns of zeros and ones will contain information closely associated with students' cognitive processes. The procedure is illustrated with basic electricity problems.

Discussion

This study displayed the successful application of the component scoring method in the domain of electricity, which is more complicated than the signed-number subtraction domain. This scoring method is applicable and useful unless the number of components is unmanageably large.

The immediate return of component scoring is being able to identify where the error was made. This leads directly to a remediation scheme for a particular error. The evaluation of component scores is difficult to do. One way to do this evaluation is to ask how objective the scoring is by examining sources of variations of scoring and the meaningfulness of these variations. Using some information similar to the table, one can answer these two questions. In this study there was only a general path that produced identical results, hence variation was trivial.

The other way to evaluate the scoring is to compare the item by item correlation matrix of component scoring and regular scoring. If the scores reflect the knowledge of electricity circuits, then the correlations of scores should be non-negative. It was found that there were eleven significantly large negative correlations for the regular scoring, while only one in the component scoring.

The number of components can be derived by first making a theoretical study of materials to be tested and further modified after the pilot study, and then validating the components by comparing the real data and the responses simulated by the erroneous rules of various components. Benefits of component scoring are far-reaching. For example, instead of a very specific conditional description needed to express the erroneous rule in one component (right or wrong responses), the component scoring method can express most of the erroneous rules by the combination of components and component-specific errors. For example if three errors are identified with each of four components then 3**4=81 combinations of possible erroneous rules exist; and each of 81 erroneous rules can be specifically defined by only 12 rules. Other benefits of the component scoring method are realized when responses are mapped on to the rule space (Tatsuoka, 1984) by using right or wrong response patterns. Some rules can have small spatial distances in the rule space but the response patterns are distinctly unique. For such a case, if the component scoring method were utilized, two response patterns can be distinguished more precisely.

2. Dowd, E. A. (1985). Computer-based education in practice: The mathematics laboratory at Urbana Junior High School. Urbana, IL: University of Illinois, Computer-based Education Research Laboratory.

Abstract

The Urbana Junior High School Mathematics laboratory, mathlab, has the PLATO system and terminals for the administration of tests and data collection for the Computerized Adaptive Testing and Measurement (CATM) research project. Various types of lessons are offered to students such as instruction, drill, tests and simulation games. The curriculum design and student record keeping capabilities of the PLATO system allow an individualized student supplement to their regular math class.

3. Eddins, J. M. (1986). A microcomputer system for diagnostic testing of cognitive errors. Proceedings of the 28th International ADCIS Conference, 279-282.

Abstract

Under sponsorship of the Office of Naval Research, Kikumi Tatsuoka at the University of Illinois has developed a measurement model called Rule Space, which can be applied to computerized tests to diagnose students' cognitive errors. Application of this technique to a training environment is now being attempted at the San Diego Naval Training Center. A flexible test administration system using microcomputers was installed, capable of administering regular as well as adaptive tests, of collecting response data, and of interacting with the Navy's CMI system. The testing system is described, and the procedures for creating and administering the special diagnostic adaptive tests are discussed.

Discussion

The mathematical model, rule space, makes it possible to generalize the adaptive testing procedure as well as map the outcome of instruction. The development of appropriate test items is basic for two reasons. First, the item parameters are a requirement for the determination of the true score (the "x" axis in the rule space concept), and the items serve as a tool for making observations and structuring the item tree.

For example, present work in teaching the use of automated graphics programs for architectural design has the same kind of item requirements even though the subject domain is different. Questions concerning the processes of representation of space and line in a computer must be developed in order to measure the user's conceptual level, in much the same way that questions were developed in fraction arithmetic. Studies of wrong approaches or inefficient methods of work can lead to a skills hierarchy or task structure. Preliminary protocol studies reveal differences in mental models for specific classes of users. The use of "shortcuts" seems to be characteristic of "experts" or those who really understand, as observations of borrowing procedures often revealed. All subject domains have similar attributes which must be discovered.

Of particular value in the rule space concept is the mapping of changes in order to descrobe and evaluate the effects of instruction. Without graphic representation, it would be difficult to draw conclusions concerning the kinds of conceptual shifts occurring with different

A Summary List of Publications

Birenbaum, M. (1986). Effect of dissimulation motivation and anxiety on response pattern appropriateness measures. Applied Psychological Measurement, 10(2), 167-174.

Birenbaum, M., & Shaw, D. (1985). Task specification chart: a key to a better understanding of test results. Journal of Educational Measurement, 22(3), 219-230.

Birenbaum, M., & Tatsuoka, K. K. (1983). The effect of scoring system based on the algorithm underlying the students' response patterns on the dimensionality of achievement test data of the problem solving type. Journal of Educational Measurement, 20(1), 17-26.

Birenbaum, M., & Tatsuoka, K. K. (1987). Effects of "on-line" test feedback, on the seriousness of subsequent errors. Journal of Educational Measurement, 24(2), 145-155

Birenbaum, M., & Tatsuoka, K. K. (in press) Open-ended versus multiple-choice response formats--it does make a difference for diagnostic purposes. Applied Psychological Measurement.

Dowd, E. A. (1985). Computer-based education in practice: The mathematics laboratory at Urbana Junior High School. Urbana, IL: University of Illinois, Computer-based Education Research Laboratory.

Eddins, J. M. (1986). A microcomputer system for diagnostic testing of cognitive errors. Proceedings of the 28th International ADCIS conference, 279-282.

Harnisch, D. L., & Tatsuoka, K. K. (1983). A comparison of appropriateness indices based on item response theory. In R.K. Hambleton, (Ed.) Applications of Item Response Theory, (pp.104-122). Vancouver, British Columbia: Educational Research Institute of British Columbia.

Shaw, D. (1986).· Cognitive error diagnosis and prescription for fraction arithmetic. Proceedings of the 28th International ADCIS conference, 324-331.

Tatsuoka, K. K. (1983). Rule space: An approach for dealing with misconceptions based on item response theory. Journal of Educational Measurement, 20(4), 34-38.

Tatsuoka, K. K. (1984a). Caution indices based on item response theory. Psychometrika, 49(1), 95-110.

Tatsuoka, K. K. (1984b). Changes in error types over learning stages.

Journal of Educational Psychology, 76(1), 120-199.

Tatsuoka, K. K. (1985a). A probabilistic model for diagnosing misconceptions in the pattern classification approach. Journal of Educational Statistics, 12(1), 55-73.

Tatsuoka, K. K. (1985b). Diagnosing cognitive errors: Statistical pattern classification and recognition approach. (Technical Report 85-1-ONR). Urbana, IL: University of Illinois, Computer-based Education Research Laboratory.

Tatsuoka, K. K. (1986a). A cognitive error diagnostic adaptive testing system. Proceedings of the 28th International ADCIS conference, 338-342.

Tatsuoka, K. K. (1986b). Diagnosing cognitive errors: Statistical pattern classification and recognition approach. Behaviormetrika, 19, 73-86.

Tatsuoka, K. K. (in press). Toward an integration of item-response theory and cognitive error diagnoses. In N. Fredericksen, R. L. Glaser, A. M. Lesgold, & M. J. Shafto (Eds.), Diagnostic Monitoring of Skill and Knowledge Acquisition. Hillsdale, NJ: Erlbaum.

Tatsuoka, K. K. (1987). Validation of cognitive sensitivity for item response curves. Journal of Educational Measurement, 24(3), 233-246.

Tatsuoka, K. K., & Eddins, J. M. (1985). Computer-based analysis of students' procedural "bugs" in arithmetic domain. Journal of Computer-based Instruction, 12(2), 34-38.

Tatsuoka, K. K., & Linn, R. L. (1983). Indices for detecting unusual patterns: Links between two general approaches and potential applications. Applied Psychological Measurement, 7(1), 81-96.

Tatsuoka, K. K., & Tatsuoka, M. M. (1983). Spotting erroneous rules of operation by the individual consistency index. Journal of Educational Measurement, 20(3), 221-230.

Tatsuoka, K. K., & Tatsuoka, M. M. (1985). Bug distribution and pattern classification (Technical Report 85-3-ONR). Urbana, IL: University of Illinois, Computer-based Education Research Laboratory.

Tatsuoka, K. K., & Tatsuoka, M. M. (1987). Bug distribution and statistical pattern classification. Psychometrika, 52(2), 193-206

Tatsuoka, K. K., & Yamamoto, K. (1985). Application of component scoring to a complicated cognitive domain (Technical Report 85-2-ONR). Urbana, IL: University of Illinois, Computer-based Education Research Laboratory.

Tatsuoka, M. M. (1986). Graph theory and its applications in educational research: A review and integration. Review of Educational Research, 56, 291-329.

Wise, S. L., & Tatsuoka, M. M. (1986). Assessing the dimensionality of dichotomous data using modified order analysis. Educational and Psychological Measurement, 46, 25-301.

END

FILMED

FEB. 1988

DTIC